

Statistics: New Foundations, Toolbox, and Machine Learning Recipes

By Vincent Granville, Ph.D.

MLTechniques.com

July 2019

This book is intended for busy professionals working with data of any kind: engineers, BI analysts, statisticians, operations research, AI and machine learning professionals, economists, data scientists, biologists, and quants, ranging from beginners to executives. In about 300 pages, it covers many new topics, offering a fresh perspective on the subject, including rules of thumb and recipes that are easy to automate or integrate in black-box systems, as well as new model-free, data-driven foundations to statistical science and predictive analytics. The approach focuses on robust techniques; it is bottom-up (from applications to theory), in contrast to the traditional top-down approach. The material is accessible to practitioners with a one-year college-level exposure to statistics and probability. The compact and tutorial style, featuring many applications with numerous illustrations, is aimed at practitioners, researchers, and executives in various quantitative fields.

New ideas, advanced topics and state-of-the-art research are discussed in simple English, without using jargon or arcane theory. It unifies topics that are usually part of different fields (machine learning, statistics, computer science), broadening the knowledge and interest of the reader in ways that are not found in any other book. This short book contains a large amount of condensed material that would typically be covered in 1,000 pages in traditional publications, including data sets, source code, and Excel spreadsheets. Thanks to cross-references and redundancy, the chapters can be read independently, in random order.

This book is based on several core articles and many tutorials that I have written over the last few years. Chapters are organized and grouped by themes: natural language processing (NLP), resampling, time series, central limit theorem, statistical tests, boosted models (ensemble methods), tricks and special topics, appendices, and so on. It is available for Data Science Central members exclusively. The text in blue consists of clickable links to provide the reader with additional references. Source code and Excel spreadsheets summarizing computations, are also accessible as hyperlinks for easy copy-and-paste or replication purposes.

About the author

Vincent Granville is a start-up entrepreneur, patent owner, author, investor, pioneering data scientist with 30 years of corporate experience in companies small and large (eBay, Microsoft, NBC, Wells Fargo, Visa, CNET) and a former VC-funded executive, with a strong academic and research background including Cambridge University.

Content

Part 1 - Machine Learning Fundamentals and NLP

We introduce a simple ensemble technique (or boosted algorithm) known as *Hidden Decision Trees*, combining robust regression with unusual decision trees, useful in the context of transaction scoring. We then describe other original and related machine learning techniques for clustering large data sets, structuring unstructured data via indexation (a natural language processing or NLP technique), and perform feature selection, with Python code and even an Excel implementation.

1. Multi-use, Robust, Pseudo Linear Regression -- page 12

- Introduction
- Example: Simulated Data with Correlated Features
- Clustering the Variables
- Clustering the Observations

2. A Simple Ensemble Method, with Case Study (NLP) -- page 15

- The Problem
- Feature Selection and Best Practices
- Methodology and Solution
- Case Study: Results
- Source Code
 - Perl, R, Python, Julia

3. Excel Implementation -- page 24

- Excel template for general machine learning
- Who should use the spreadsheet?
- Description of the techniques used
- Spreadsheet versus Python version
- Why a brand new set of machine learning tools?
- The Spreadsheet
- Confidence intervals for the response

4. Fast Feature Selection -- page 31

- Predictive Power of a Feature, Cross-Validation
- Data structure, computations

5. Fast Unsupervised Clustering for Big Data (NLP) -- page 36

- Building a Keyword Taxonomy
- Fast Clustering Algorithm
- Computational Complexity

6. Structuring Unstructured Data -- page 40

- Indexation algorithm
 - Potential improvement
-

Part 2 - Applied Probability and Statistical Science

We discuss traditional statistical tests to detect departure from randomness (the null hypothesis) with applications to sequences (the observations) that behave like stochastic processes. The central limit theorem (CLT) is revisited and generalized with applications to time series (both univariate and multivariate) and Brownian motions. We discuss how weighted sums of random variables and stable distributions are related to the CLT, and then explore mixture models -- a better framework to represent a rich class of phenomena. Applications are numerous, including optimum binning for instance. The last chapter summarizes many of the statistical tests used earlier.

7. Testing for Randomness -- page 42

- Context
- Methodology
 - Algorithm to compute the observed gap distribution
 - Statistical testing
- Application to Number Theory Problem
 - A counter-example
 - Potential use in cryptography
- Conclusion

8. The Central Limit Theorem Revisited -- page 48

- A special case of the Central Limit Theorem
- Simulations, testing, and conclusions
 - The Lyapunov connection
- Generalizations
 - Correlated observations
 - Non-random (deterministic) observations
 - Other generalizations
- Source code

9. More Tests of Randomness -- page 55

- Central Limit Theorem for Non-Random Variables
- Testing Randomness: Max Gap, Auto-Correlations and More
 - Convergence to a non-degenerate distribution
- Excel Spreadsheet with Computations
- Potential Research Areas
 - Generalization to higher dimensions

10. Random Weighted Sums and Stable Distributions -- page 63

- Central Limit Theorem: New Approach
 - Theorem
- Stable and Attractor Distributions
 - Using decaying weights
 - Exact distribution
 - More about stable distributions and their applications
- Non CLT-compliant Weighted Sums, and their Attractors
 - Testing for normality
 - Testing for symmetry and dependence on kernel
 - Testing for uni-modality and other peculiarities
 - Testing for semi-stability
- Conclusions

11. Mixture Models, Optimum Binning and Deep Learning -- page 73

- Introduction and Context
- Approximations Using Mixture Models
 - The error term
 - Kernels and model parameters
 - Algorithms to find the optimum parameters
 - Convergence and uniqueness of solution
 - Find near-optimum with fast, black-box step-wise algorithm
- Example
 - Data and source code
 - Results
- Applications
 - Optimal binning
 - Predictive analytics
 - Test of hypothesis and confidence intervals
 - Deep learning: Bayesian decision trees
 - Clustering
- Interesting problems
 - Gaussian mixtures uniquely characterize a broad class of distributions
 - Weighted sums fail to achieve what mixture models do
 - Stable mixtures

- Nested mixtures and Hierarchical Bayesian Systems
- Correlations

12. Long Range Correlations in Time Series -- page 87

- Introduction and time series deconstruction
 - Example
 - Deconstructing time series
 - Correlations, Fractional Brownian motions
- Smoothness, Hurst exponent, and Brownian test
 - Our Brownian tests of hypothesis
 - Data
- Results and conclusions
 - Charts and interpretation
 - Conclusions

13. Stochastic Number Theory and Multivariate Time Series -- page 95

- Some Definitions
- Digits Distribution in b -processes
- Strange Facts and Conjectures about the Rabbit Constant
- Gaming Application
 - De-correlating Time Series Using Mapping and Thinning Techniques
 - Dissolving the Auto-correlation Structure Using Multivariate b -processes

14. Statistical Tests: Summary -- page 101

- General Methodology
- Off-the-beaten-path Statistical Tests

Part 3 - New Foundations of Statistical Science

We set the foundations for a new type of statistical methodology fit for modern machine learning problems, based on generalized resampling. Applications are numerous, ranging from optimizing cross-validation to computing confidence intervals, without using classic statistical theory, p -values, or probability distributions. Yet we introduce a few new fundamental theorems, including one regarding the asymptotic properties of generic, model-free confidence intervals.

15. Modern Resampling Techniques for Machine Learning -- page 107

- Re-sampling and Statistical Inference
 - Main Result
 - Sampling with or without Replacement

- Illustration
 - Optimum Sample Size
 - Optimum K in K -fold Cross-Validation
 - Confidence Intervals, Tests of Hypotheses
- Generic, All-purposes Algorithm
 - Re-sampling Algorithm with Source Code
 - Alternative Algorithm
 - Using a Good Random Number Generator
- Applications
 - A Challenging Data Set
 - Results and Excel Spreadsheet
 - A New Fundamental Statistics Theorem
 - Some Statistical Magic
 - How does this work?
 - Does this contradict entropy principles?
- Conclusions

16. Model-free, Assumption-free Confidence Intervals -- page 121

- Principle
- 2. Examples
 - Estimator used in nearest neighbors clustering
 - Weighted averages when dealing with outliers
 - Correlation coefficient estimated via re-sampling
 - Auto-correlated time series, U -statistics
- Counterexamples
- Estimating A
- Estimating B
 - Getting more accurate values
 - Getting even more accurate values
- Theoretical Background
 - Connection with the re-scaled range and the Hurst exponent
 - General case
 - Another approach to building confidence intervals
- Conclusions

17. The Distribution of the Range: A Beautiful Probability Theorem -- page 133

- Theorem and proof
- Connection with order statistics and the Renyi Representation

Part 4 - Case Studies, Business Applications

These chapters deal with real life business applications. Chapter 18 is peculiar in the sense that it features a very original business application (in gaming) described in details with all its components, based on the material from the previous chapters. Then we move to more traditional machine learning use cases. Emphasis is on providing sound business advice to data science managers and executives, by showing how data science can be successfully leveraged to solve problems. The presentation style is compact, focusing on strategy rather than technicalities.

18. Gaming Platform Rooted in Machine Learning and Deep Math -- page 136

- Description, Main Features and Advantages
- How it Works: the Secret Sauce
 - Public Algorithm
 - The Winning Numbers
 - Using Seeds to Find the Winning Numbers
 - ROI Tables
- Business Model and Applications
 - Managing the Money Flow
 - Virtual Currency
- Challenge and Statistical Results
 - Data Science / Math Competition
 - Controlling the Variance of the Portfolio
 - Probability of Cracking the System
- Designing 16-bit and 32-bit Systems
 - Layered ROI Tables
 - Smooth ROI Tables
 - Systems with Winning Numbers in $[0, 1]$

19. Digital Media: Decay-adjusted Rankings -- page 148

- Introduction
- Top DSC blogs
- Interesting Insights
- New Scoring Engine
- Good versus perfect model
- Next steps

20. Building a Website Taxonomy -- page 153

- Seed Keywords
- General Methodology
- Top 2,500 Data Science Websites

- Data and Source Code
- Detailed Methodology
- Possible Improvements

21. Predicting Home Values -- page 158

- The data
- Leveraging available data, getting additional data
- Potential metrics to consider
- Model selection and performance

22. Growth Hacking -- page 161

- Growth Hacking: Part I
 - Strategy
 - Methodology
 - Scoring algorithm
 - Data Sets, Excel spreadsheet
 - Python Source Code
 - Next steps
- Growth Hacking: Part II
- Growth Hacking: Part III
 - Algorithm: categorizing / clustering articles
- Conclusions

23. Time Series and Growth Modeling -- page 169

- Case Study: The Problem
 - Business questions
- Deep Analytical Thinking
 - Answering hidden questions
- Data Science Wizardry
 - Generic algorithm
 - Illustration with three different models
 - Results
- A few data science hacks

24. Improving Facebook and Google Algorithms -- page 179

- Five Case Studies
 - More about the Facebook ad processing system
- Why so many Machine Learning Implementations Fail?
 - The fake news issue
 - When machine learning is used as a scapegoat
- Twenty four tips for better data science

Part 5 - Additional Topics

Here we cover a large number of topics, including sample size problems, automated exploratory data analysis, extreme events, outliers, detecting the number of clusters, p -values, random walks, scale-invariant methods, feature selection, growth models, visualizations, density estimation, Markov chains, A/B testing, polynomial regression, strong correlation and causation, stochastic geometry, K nearest neighbors, and even the exact value of an intriguing integral computed using statistical science, just to name a few.

25. Solving Common Machine Learning Challenges -- page 187

- Eliminating sample size effects
- Sample size determination
- Automatically detecting the number of clusters
- Fixing issues in regression models
- Performing joins on mismatched data
- Scale invariant techniques
- Blending data sets with non-compatible fields
- Automated exploratory data analysis
- Simple solution to feature selection problems
- Coefficient of Correlation for Non-Linear Relationships
- Choosing a regression model
- Growth modeling with Excel
- Interesting charts
- Simplified logistic regression

26. Outlier-resistant Techniques, Cluster Simulation, Contour Plots -- page 214

- General Framework
 - Finding a robust centroid
 - Generalization to linear regression problems
 - General outlier detection techniques
 - A related physics problem
- Algorithm to find centroid when $p > 1$
 - Source code to generate points and compute centroid
 - Generating point clouds with Monte Carlo simulation
- Examples and results
- Convergence of the algorithm
- Interesting Contour Maps

27. Strong Correlation Metric -- page 225

- Definition of strong correlation

- Comparison with traditional (weak) correlation
- Excel spreadsheet with computations and examples
- When to use strong versus weak correlation?
- Generalization

28. Special Topics -- page 229

- Comparing ML, Data Science, AI, Deep Learning, and Statistics
 - Different Types of Data Scientists
 - Machine Learning versus Deep Learning
 - Machine Learning versus Statistics
 - Data Science versus Machine Learning
- Distribution of Arrival Times for Extreme Events
 - Simulations
 - Theoretical Distribution of Records over Time
 - Useful Results
- How to Lie with p Values?
- Off-the-beaten-path Machine Learning Topics
 - Random walks in one, two and three dimensions
 - Estimation of the convex hull of a set of points
 - Constrained linear regression on unusual domains
 - Robust and scale-invariant variances
 - The Tweedie distributions
 - The arithmetic-geometric mean
 - Weighted version of the K -NN clustering algorithm
 - Multivariate exponential distribution and storm modeling
- Variance, Clustering, and Density Estimation Revisited
 - Working on the Grid, not on the Original Space
 - Density Estimation
 - Supervised Clustering
 - Scale-Invariant Variance
 - Historical Notes
- New K -NN Clustering Algorithm and Data Reduction
- Spatial Patterns Found in Random Points
- Stochastic Geometry: Spatial Coverage Problem
- Markov Chains and the Collatz Conjecture
- Special Integral Solved Using Statistical Concepts
- From A/B Testing to Discrete Choice Analysis
- Deep Dive into Polynomial Regression and Overfitting
- Lifecycle of Data Science Projects

Appendix A. Linear Algebra Revisited -- page 266

- Power of a Matrix

- Examples, Generalization, and Matrix Inversion
 - Example with a non-invertible matrix
 - Fast computations
- Application to Machine Learning Problems
 - Markov chains
 - Time series
 - Linear regression
- Appendix

Appendix B. Stochastic Processes and Organized Chaos -- page 272

- General framework, notations and terminology
 - Finding the equilibrium distribution
 - Auto-correlation and spectral analysis
 - Ergodicity, convergence, and attractors
 - Space state, time state, and Markov chain approximations
 - Examples
- Case study
- Applications
- Additional topics
 - Perfect stochastic processes and Brownian motions
 - Characterization of equilibrium distributions (the attractors)
 - Probabilistic calculus, number theory, special integrals
- Appendix
 - Computing the autocorrelation at equilibrium
 - Proof of the first fundamental theorem
 - How to find the exact equilibrium distribution
 - Perfect process with no autocorrelation

Appendix C. Machine Learning and Data Science Cheat Sheet -- page 297

- Hardware
- Linux environment on Windows laptop
- Basic UNIX commands
- Scripting languages
- Python, R, Hadoop, SQL, DataViz
- Machine Learning
 - Algorithms
 - Getting started
 - Applications
- Data sets and sample projects